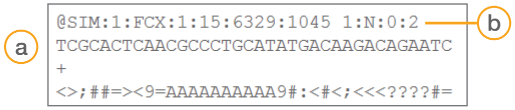
Analyzing Data Using the Enrichment Workflow

Analyze

The Enrichment workflow aligns reads against the whole genome reference and performs variant analysis for regions specified in a manifest file.

1 FASTQ File Generation

- Per-sample FASTQ files are generated by matching index reads to sample sheet indices, and then saved in gzipped format. FASTQ files contain only reads passing filter (PF) and their base call quality scores.
- FASTQ files are written to the MiSeqAnalysis folder in **Data\Intensities\BaseCalls** and then copied to the MiSeqOutput folder.
- File name—**samplename_sample#_lane#_read#_001.fastq.gz**.
- With each analysis, FASTQ files are overwritten. REQUEUE
- Reads are randomly subsampled to produce Assemble_N_Rx.fastq.gz files, where N is the order of the sample number and x is the read number, that contain reads used in the assembly process.



```
@SIM:1:FCX:1:15:6329:1045 1:N:0:2
TCGCACTCAACGCCCTGCATATGACAAGACAGAATC
+
<>;##=><9=AAAAAAAAAA9#:<#<;<<<????#=
```

a FASTQ is a text-based file format with four lines per read: identifier, sequence, plus sign (+), and quality score in ASCII format.

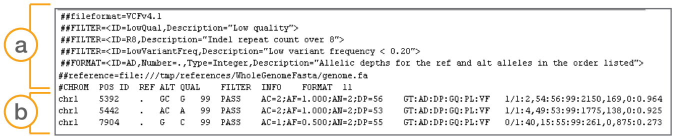
b FASTQ file identifier format: @Instrument:RunID:FlowCellID:Lane:Tile:X:Y ReadNum:FilterFlag:0:SampleNumber

2 Alignment

- BWA aligns reads against the reference genomes specified in the sample sheet.
- Automatically adjusts parameters based on read lengths and error rates, and then estimates insert size distribution.
- Generates an industry standard BAM (*.bam) file (binary version of a SAM file) that contains sequence alignment data.
- BAM files are written to the MiSeqAnalysis folder in **Data\Intensities\BaseCalls\Alignment** and then copied to the MiSeqOutput folder.
- File name—**samplename_sample#.bam**, where # is the order samples are listed in the sample sheet.
- BAM files must be viewed using a Genome Browser, such as Broad's Integrated Genome Viewer (IGV).
- With each analysis, new BAM files are generated in a sequentially numbered Alignment folder. REQUEUE

3 Variant Calling

- GATK identifies SNPs and short indels, and generates a Variant Call Format (*.vcf) file.
- (Optional alternative) The Somatic Variant Caller identifies variants present at low frequency and minimizes false positives, and generates a VCF file.
- VCF files are written to the MiSeqAnalysis folder in **Data\Intensities\BaseCalls\Alignment** and then copied to the MiSeqOutput folder.
- File name—**samplename_sample#.vcf**, where # is the order samples are listed in the sample sheet.
- With each analysis, new VCF files are generated in a sequentially numbered Alignment folder. REQUEUE



```
##fileformat=VCFv4.1
##FILTER=<ID=LowQual,Description="Low quality">
##FILTER=<ID=R8,Description="Indel repeat count over 8">
##FILTER=<ID=LowVariantFreq,Description="Low variant frequency < 0.20">
##FORMAT=<ID=AD,Number=.,Type=Integer,Description="Allelic depths for the ref and alt alleles in the order listed">
##reference=file:///tmp/references/WholeGenomeFasta/genome.fa
#CHROM  POS ID  REF ALT QUAL    FILTER  INFO    FORMAT  11
chr1    5392    .   GC  G   99  PASS    AC=2;AF=1.000;AN=2;DP=56    GT:AD:DP:GQ:PL:VF   1/1:2,54:56:99:2150,169,0:0.964
chr1    5442    .   AC  A   99  PASS    AC=2;AF=1.000;AN=2;DP=53    GT:AD:DP:GQ:PL:VF   1/1:4,49:53:99:1775,138,0:0.925
chr1    7904    .   G   C   99  PASS    AC=1;AF=0.500;AN=2;DP=55    GT:AD:DP:GQ:PL:VF   0/1:40,15:55:99:261,0,875:0.273
```

a VCF file header—Lists VCF annotations used in the file and includes column headings for the data lines.

b Data lines—Each line contains information about a single variant.

Optional Sample Sheet Settings

See the MiSeq Sample Sheet Quick Reference Guide for values and other settings.

- **Adapter**—Specify the adapter sequence to prevent reporting beyond sample DNA.
- **VariantCaller**—Specify Somatic Variant Caller or Starling as an alternative to GATK.
- **FlagPCRDuplicates**—By default, excludes PCR duplicates from variant calling.
- **QualityScoreTrim**—Trims 3' ends of non-indexed reads with low quality scores. BWA default value is 15.

Optional Configurables in MiSeq Reporter.exe

For optional values, see the MiSeq Reporter User Guide

- **CreateFASTQForIndexReads**—Set to 1 to generate FASTQ files for index reads.
- **FilterNonPFReads**—Set to 0 to include non-PF reads in FASTQ files.

Documentation

- MiSeq Reporter Assembly Workflow Reference Guide
- MiSeq Reporter Theory of Operation
- MiSeq Sample Sheet Quick Ref Guide
- MiSeq Run Folder Quick Ref Card

Additional Information

- SAM Tools—http://samtools.sourceforge.net/
- BWA—http://bio-bwa.sourceforge.net/
- GATK—http://www.broadinstitute.org/gatk/index.php
- IGV—http://www.broadinstitute.org/software/igv/
- VCF Tools—http://vcftools.sourceforge.net/